네트워크 모델을 이용한 새로운 회귀분석방법

김기복, 인치호, 김희석 *
세명대학교 컴퓨터과학과, 청주대학교 전자공학과 *

A new regression analysis method in network model

Gi-Bog Kim, Hi-Suk Kim, Chi-Ho Lin
Dept. of Computer Science, Semyung University
E-mail: kgb1587@hanmail.net,

Abstract

본 논문에서는 네트워크가 막연히 무작위적이라고 하기에는 사회나 세포, 인터넷 등이 어떤 법칙에 따라 짜연진 것처럼 보인다. 하지만 복잡한 네트워크의 모습이 네트워크의 모델과 실제로 똑같은지를 비교하기는 그리 쉬운 문제가 아니다. 무작위적 네트워크의 경우는 수학적으로 엄밀히 말하자면 쁘아송분포를 따른다. 쁘아송분포에서는 모든 점들이 동일한 확률로 여러 점들에 연결되는 기회를 갖는다. 즉 균일한 분포이다. 따라서 상당히 적거나 반대로 상당히 많은 수의 연결선을 가진 점은 극히 드물다. 이 경우 연결선 분포가 종 모양이 된다. 대부분의 점들이 곡선에 해당하는 연결선 수를 갖게 된다.

본 논문에서 쁘아송분포와 회귀분석을 통하여 하나 또는 둘 이상의 변수들 사이에 어떤 관계를 함수관계로 나타내어 분석하는 방법을 보이고 회귀분석 방법에 의해서 미래를 예측하고자 한다.

I. 서론

회귀분석은 여러 변수들 사이의 관계를 알아보고자 하는 경우에 많이 사용되는 분석법이다.[1]
물론 두 변수 사이의 관계를 알아보기 위해서 가장 먼저 하는일은 산점도를 그려보거나 변수사이의 관계를 나타내는 상관 관계를 구해보는 것이다. 그러나 상관분석은 두 변수 사이에 직선 관계가 있는지의 여부 만을 알아보는 것이 목적이며 두 변수 사이에 구체적으로 어떠한 함수 관계가 있는 가를 말해주지는 않는다. 따라서 변수들 사이의 관계를 규명한 후 이를 이용하여 한 변수의 값으로부터 다른 변수를 예측하고자 하는 경우에는 그 관계를 함수식으로 나타내는 것이 필요하다.[2]
본 논문에서는 이러한 예측을 하기위해서 SAS 프로그램을 이용하여 보다 효율적으로 구현하고자 한다.

II SAS를 이용한 분석

2 .1 Poisson distribution

확률함수 ranpoi(seed, λ)는 평균 발생횟수가 λ인 포아송확률변수의 관측치를 발생

```
data rpois;
do i=0 to 1000;
rp = ranpoi(90589, 20);
output; end;
proc print;
run;
```

시간당 평균발생건수가 20인 포아송확률변수로부터 시간당 발생건수가 0부터 100인 경우까지의 확률분포함수와 확률질량함수를 구하고 그래프로 알아보자(실제, 발생건수는 0부터 ∞까지 확률값이 존재한다.)

```
data pois;
do x=0 to 100;
p= poisson(20, x);
m=poisson(20, x) - poisson(20, x - 1);
output;  end;
symbol1 c= red v= none i=needle;
symbol2 c=green v= none i=needle;
proc print;
run;
proc gplot;
plot p*x=1;
plot m*x=2;
run;
```

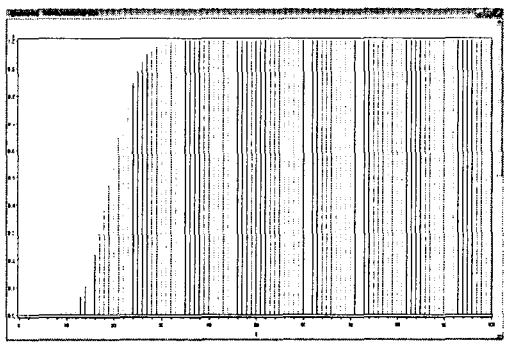

그림 1 쁘아송 분포P(20)의 확률분포함수

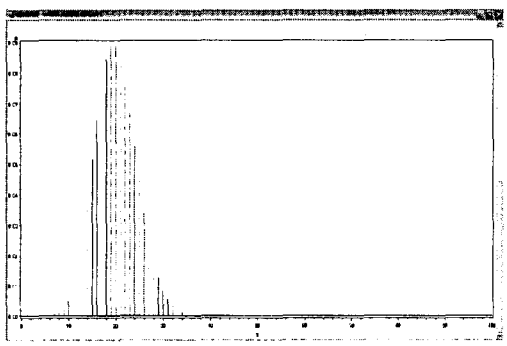

그림 2 쁘아송 분포P(20)의 확률질량함수

2.2 회귀분석

단순회귀분석은 하나의 독립변수와 이에 대응되는 평균 반응변수 사이의 선형관계를 가정한다. 즉

$Y=\alpha+\beta x+\varepsilon,\ \varepsilon \sim N(o,\ \sigma^2)$ 이며

이를 모회귀모형(population regression model)이라 부른다.

x 가 주어진 경우의 조건부 기댓값은

$E[Y|x]=\alpha+\beta x$ 인 x에 대한 일차함수로 주어지는데, 이식을 모회귀직선 (population regression line) 이라 한다.

2 .3 모회귀 모형의 추정

$x_1, x_2, x_3 \cdots, x_n$는 독립변수(설명변수)이고, $Y_1, Y_2, Y_3 \cdots, Y_n$는 종속변수(반응변수)이며, $Y_i=\alpha+\beta x_i+\varepsilon_i\ (i=1,2,\cdots)$ 을 고려하자, 여기서 ε_i 는 오차항이며, j로 독립인 $N(0,\sigma^2)$이다.

이제 $SS=\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)^2$ 을 최소로 하는 α, β를 구하자, α, 와 β에 대한 편도함수는

$\frac{\partial SS}{\partial \alpha}=-2\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)$

$\frac{\partial SS}{\partial \beta}=-2\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)$이며 이 값을 0으로 하면 $\sum_{i=1}^{n}Y_i=n\alpha+\beta\sum_{i=1}^{n}x_i$

$\sum_{i=1}^{n}x_iY_i=\alpha\sum_{i=1}^{n}x_i+\beta\sum_{i=1}^{n}x_i^2$ 를 얻는다. 그러므로 SS를 최소로 하는 α, 와 β값은 $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{x}$, $\hat{\beta}=\frac{\Sigma x_iY_i-n\overline{xY}}{\Sigma x_i^2-n\overline{x}^2}$ 이며 $\hat{\alpha}$ 와 $\hat{\beta}$를 α, 와 β의 최소 제곱추정량(Least squares estimator)이라 하며 $\hat{Y}=\hat{\alpha}+\hat{\beta}x$ 을 추정회귀직선(estimated regression line) 혹은

표본회귀직선(simple regression line)이라 한다.

2 .4 단순선형회귀 모형

단순선형회귀모형

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i,\ i=1,2,\cdots$$

β_0과β_1는 회귀계수이며 ε_i는 오차항으로서 서로 독립인 $N(0,\ \sigma^2)$확률변수

단순선형회귀모형에서 회귀계수인 β_0와β_1의 추정량은 오차의 제곱합을 최소로 하는 최소 제곱법에 의해

구할 수 있으며 이에 의해 구해지는 β_0 과 β_1를 최소제곱추정량(least squares estimator)이라 한다.

2.5 다중선형회귀분석

다중회귀 분석에서는 행렬과 벡터를 이용하면, 식의 표현과 계산작업을 쉽게 할수 있다.
따라서 행렬과 벡터를 사용하여 위의 다중선형 회귀모형을 나타내 보면 다음과 같다.

$$y = X\beta + \varepsilon$$

여기서

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \beta=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

으로 각각 정의 된다.

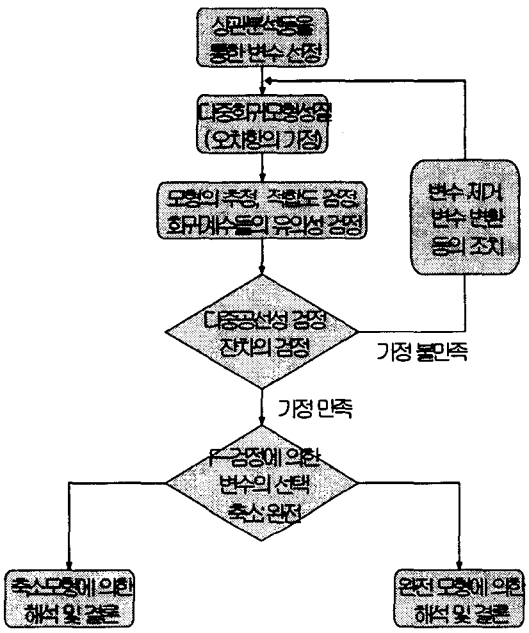

그림 3 다중선형 회귀분석의 절차

Ⅲ. SAS를 이용한 회귀분석

3.1 선형회귀분석

```
data aaa;
input x y @@;
cards;
1 50 1 45 2 31 2 30 3 28 3 25 3 23 4 25 4 22 5 21
;
proc plot data=aaa;
     plot y*x= '+'/vpos=20;
proc reg data = aaa ;
          model y=x/dw;
           output out = aaaout1 p=pred r=resid
          student=stud;
proc plot data=aaaaout1;
          plot y*x ='0'
          pred*x = 'p' / overlay vpos = 20 ;
proc plot data =aaaaout1 hpercent = 50
           vpercent = 50 ;
           plot (resid stud)* x/vref=0;
            plot (resid stud)* pred/vref=0;
proc univariate data = aaaout normal plot;
          var resid;
run;
```

결과

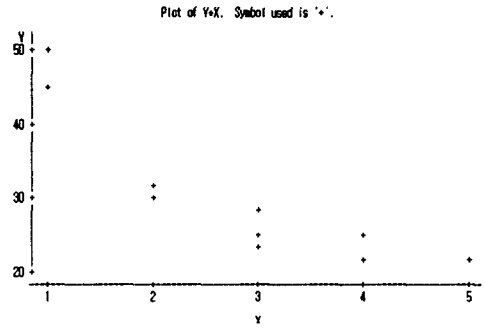

그림 4 기술적 통계량

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	666.92308	666.92308	25.77	0.0010
Error	8	207.07692	25.88462		
Corrected Total	9	874.00000			

Root MSE	5.08769	R-Square	0.7631
Dependent Mean	30.00000	Adj R-Sq	0.7335
Coeff Var	16.95898		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	48.30769	3.94932	12.23	<.0001
X	1	-6.53846	1.28813	-5.08	0.0010

그림 5 회귀모형의 적합도 검정

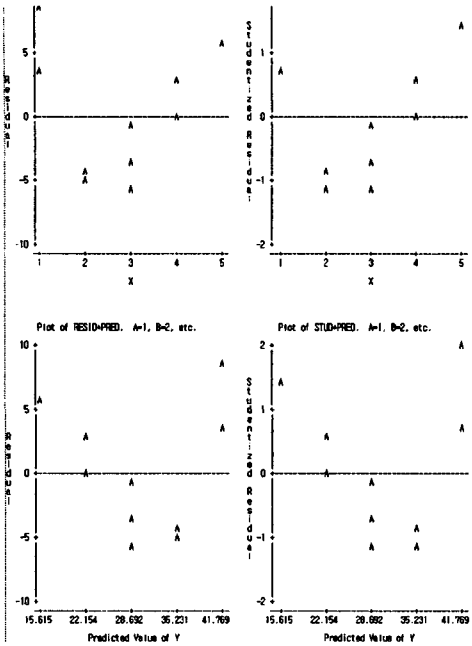

그림 6 오차항의 등분산성에대한 가정이 만족되는지 여부를 판단하기 위한 여러 가지 산점도

Ⅳ결론

본 논문에서는 현대의 우리들은 과학문명의 발달과 쏟아지는 정보속에서 정보를 효율적으로 수집하고, 정리 요약하거나 제한된 정보를 이용하여 불확실성에 대한 예측과 의사결정을 위한 절대적인 도구 역할을 하는 통계학의 이론과 SAS프로그램을 이용하여 선형회귀분석과 다중회귀분석에대한 이해와 분석방법을 통하여 미래예측에 보다 나은 방법을 제시하고 있다.

References

1) Statistics Datum Analysis, Freedom Academy, 1998

2) Survey on Statistics, Freedom Academy 1995

3) Introduction to Statistics, Kyungmun-sa 1997

4) SAS Statistical Analysis, Kyowoo-sa

5) Introduction to probability and statistics, J.S.Milton and J.C.Arnold 1990

6) SAS/STAT User's Guide, Version 6, 4th edition, 1990

7) SAS/GRAPH Software, Version 6, 1st edition,

1990